STRUCTURE OF THE ANALYSIS

The metrics previously computed (Eccentricity, .. ) have been using as features to predict the various outcomes (Endoscopic, Clinical, .. ). We used – specularly as before – a ten-fold cross validation structure in which in 9 fold we do cross validation and in the test fold we evaluate our algorithm. Due to time constraint (we have 10 metric for each cohort), we focused on SVM with radial kernel and Random forest. In random forest, we selected the top20 feature for the variance importance in each cohort, in each step of the cross validation, and we use those as features. Finally, we catered all the prediction together and we analyzed which genus, family or taxa are appearing more.

MISSINGESS

Due to the structure of the metrics, there are some missing values in the various cohorts. More refined techniques exist, and I am open to suggestion, but for the moment I have used 2 main missing techniques.

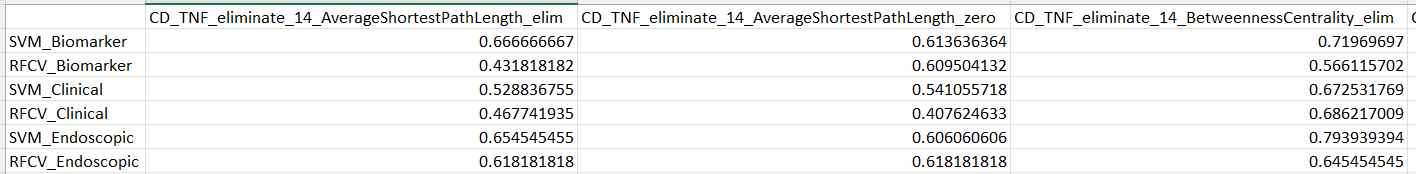
1. *Zero* 🡪 I substitute the missing value with 0, meaning that we do not know anything for that particular metric
2. *Elim* 🡪 I erase every node in which there is at least one missing node.

Hence, an instance of the prediction called “***CD\_TNF\_14\_AverageShortestPathLength\_elim***” mean that the analysis was done on CD\_TNF ; at timepoint w14; on the AverageShortestPathLength metric, and the missing were treated through elimination.

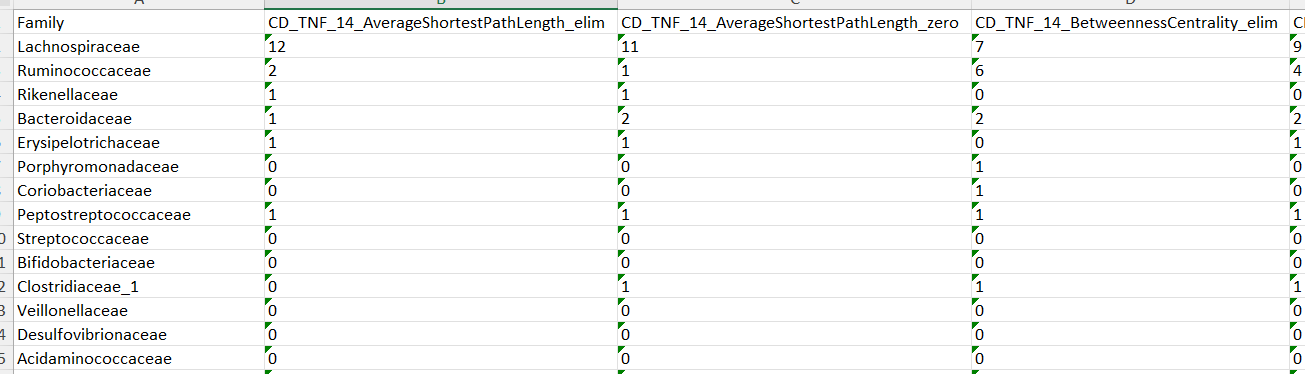
FILES

Two main folders:

1. Prediction folder: Here we catered together the prediction AUC for evey instance: i.e., for every cohort, every timepoint, every metric, every missinges treatment and every outcome (Clinical, .. ). I have also produced the same prediction but divided per metric, to make the exploration easy. The structure is the same as the previous analysis on the edge weights. All files are available as excel sheet and as R objects.



1. Important\_taxas: Here we collected various information about which taxa are important (i.e., which nodes). In the subfolders Important\_metrics/RFCV or Important\_metrics/SVM there is the ranked list of important taxa in prediction for the algorithm. Those taxa are aggregated at different levels: Taxa, Genus and Family and for the 3 external outcome (Biomarkers, Clinical, Endoscopic).



As an example,

Here we can see the most present families. Each cohort appears twice: once for each NA treatment method.